Department of Civil Engineering
National Institute of Technology, Tiruchirappalli

Landslide Susceptibility Mapping for Nilgiris District using Advanced Machine Learning Methods

By:

Sri Raaghav M D 103122060

Table of Content

Abstract

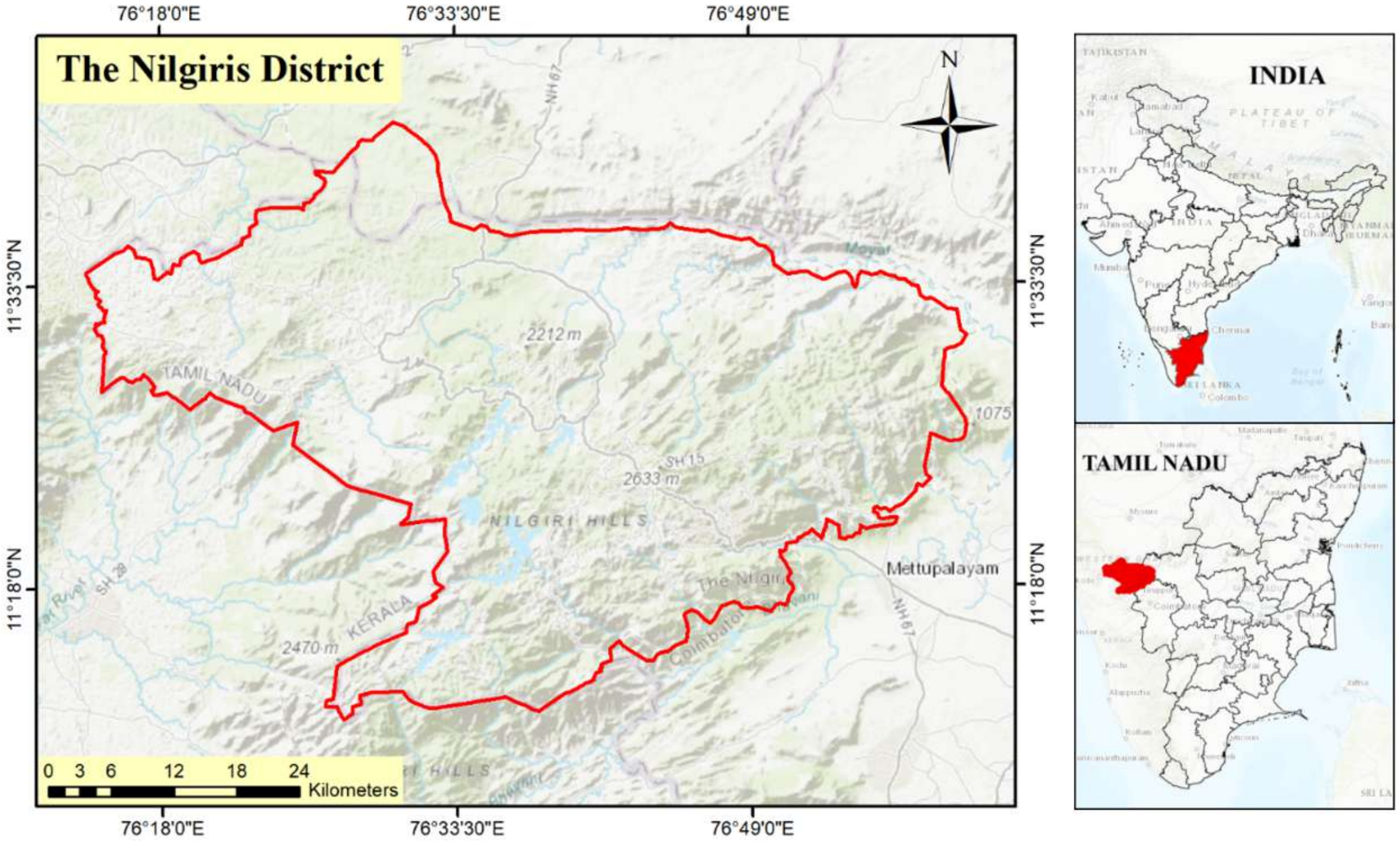
- Landslides are one of the major natural disasters in India, and this study focuses on improving landslide susceptibility mapping for the Nilgiris district in Tamil Nadu using GIS and machine learning.
- The study used many factors from geospatial databases, including elevation, slope, soil, lithology, land use, and distances to roads, streams, and lineaments, along with several terrain indices.
- Six machine learning models namely, logistic regression, CatBoost, XGBoost, support vector machine, random forest, and C5.0 were trained using past landslide data from the region.
- These models were combined into an ensemble through stacking, and validation using the ROC curve showed a high prediction accuracy with an AUC of 0.98, mapping landslide risk from very low to very high.

Introduction

- Landslides are the downward movement of rocks, soil, or other slope materials, and in India they often occur because of heavy rainfall, urbanization, and deforestation.
- Four regions in India are highly prone to landslides namely, the Himalayas, Western Ghats, Eastern Ghats, and Andaman and Nicobar Islands. The Nilgiris in the Western Ghats is one of the most affected.
- The Nilgiris district is known for its natural beauty and pleasant weather, attracting tourists throughout the year and leading to rapid urban growth from tourism income.
- The combination of steep slopes, intense rainfall, and human activities such as road widening, deforestation, and unplanned construction has made the area more unstable.
- This study uses a combined model of six algorithms trained on Bhukosh and Bhuvan data to predict landslide-prone zones accurately and support safer land-use and disaster management.

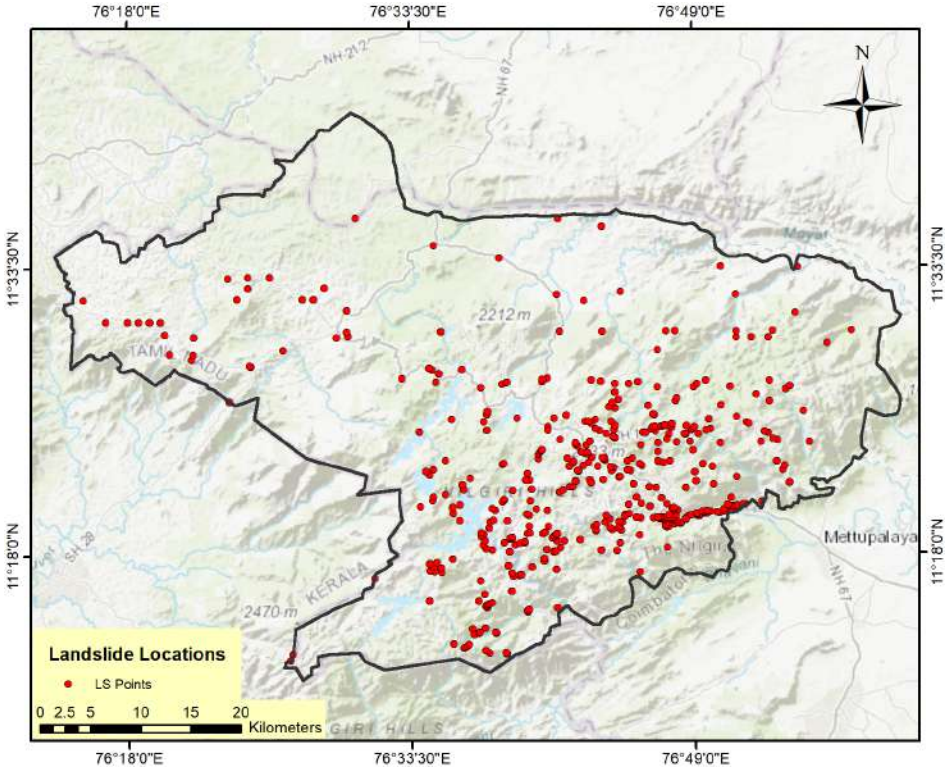
Study Area Description

- The Nilgiris district in Tamil Nadu is located at 11.4916° N and 76.7337° E and forms part of the Western Ghats.
- It covers 2,545 sq km, has a population of 7,35,394 as per the 2011 census, and includes Doddabetta, the highest peak in South India at 2,595 m.
- In 2020, about 59% of the district was covered by natural forest, and the cultivated area in 2019–2020 was 73,406 ha.
- Horticulture is practiced in three seasons, with temperate crops like tea and potato at higher altitudes, coffee and oranges at mid-levels, and spices and tropical fruits at lower levels.
- The district has mild weather with temperatures between 19.3 °C and 29.2 °C, receives 1,640.2 mm of rainfall, and attracts tourists year-round.

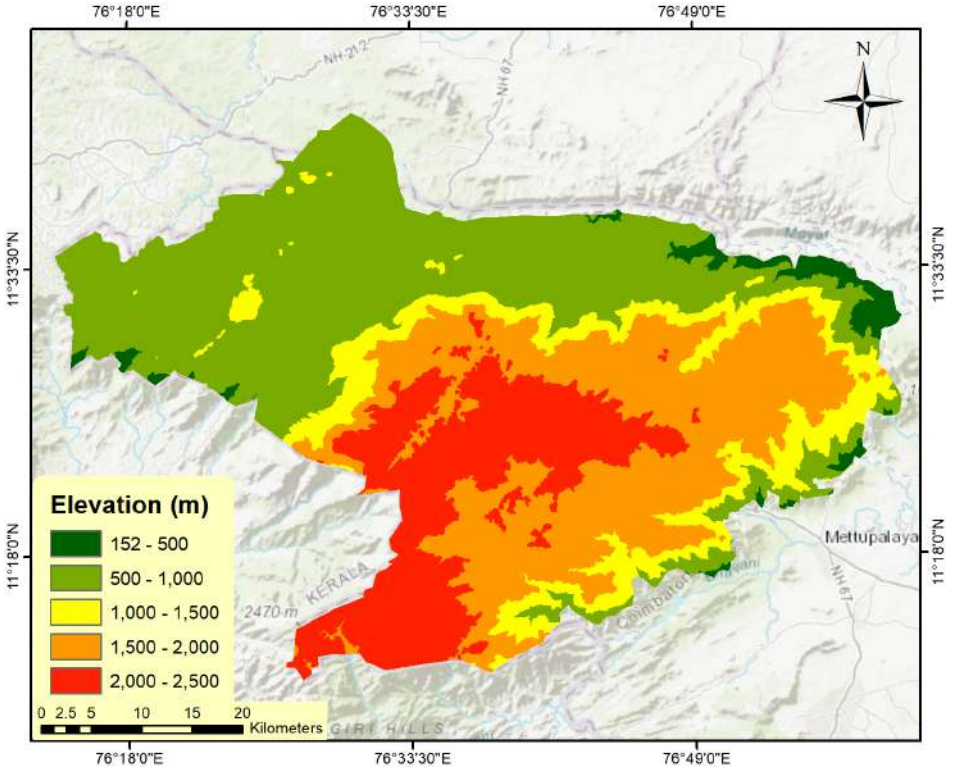

The Nilgiris District
76°18'0"E
76°33'30"E
76°49'0"E
11°33'30"N
11°18'0"N
N
TAMIL NADU
KERALA
NILGIRI HILLS
2212 m
2633 m
2470 m
1075
Mettupalayam
Moyar
SH 15
SH 28
NH 67
0 3 6 12 18 24
Kilometers
INDIA
TAMIL NADU

Landslide Factors Dataset

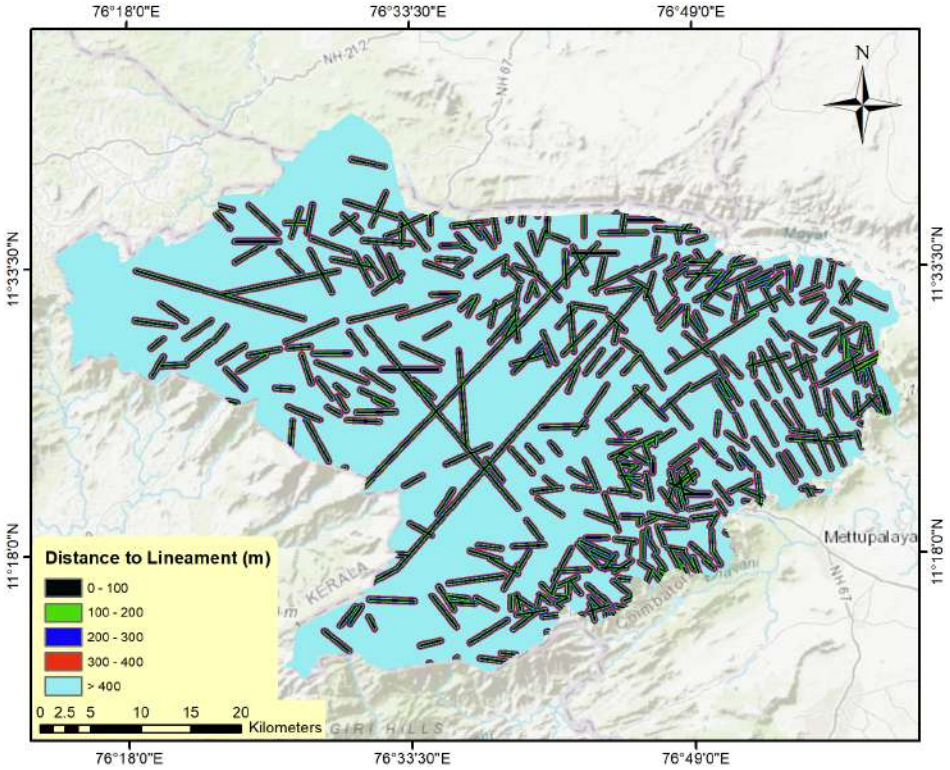
Landslide Locations – Historical landslide occurrences used for training the predictive model.



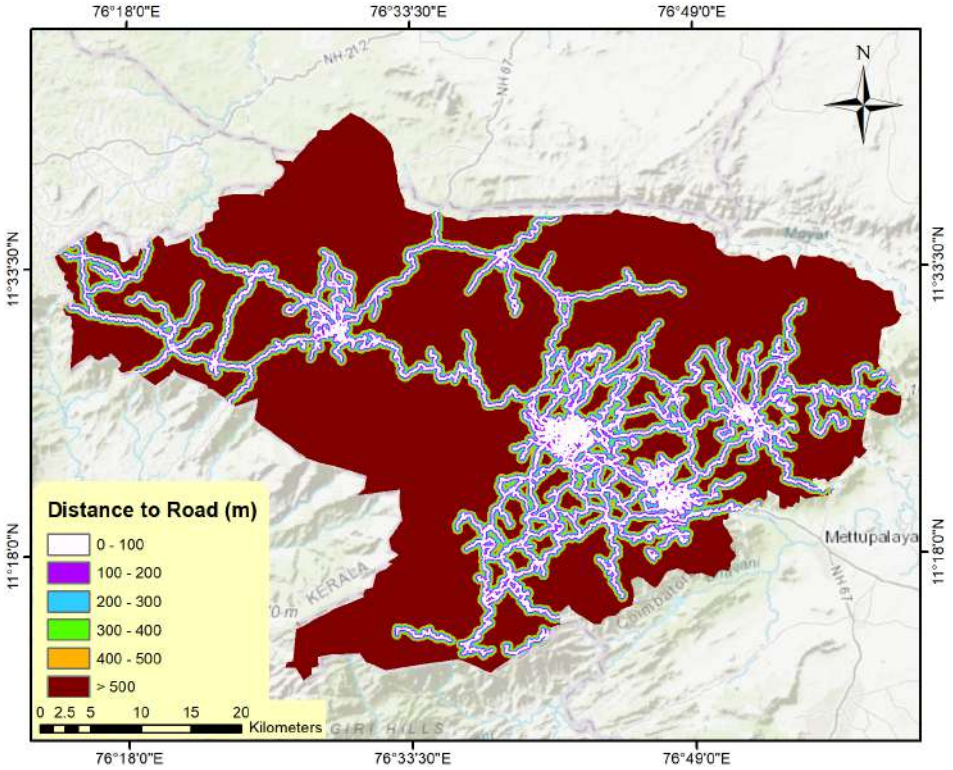
Elevation – Height above sea level, affecting climate, vegetation, and slope conditions.



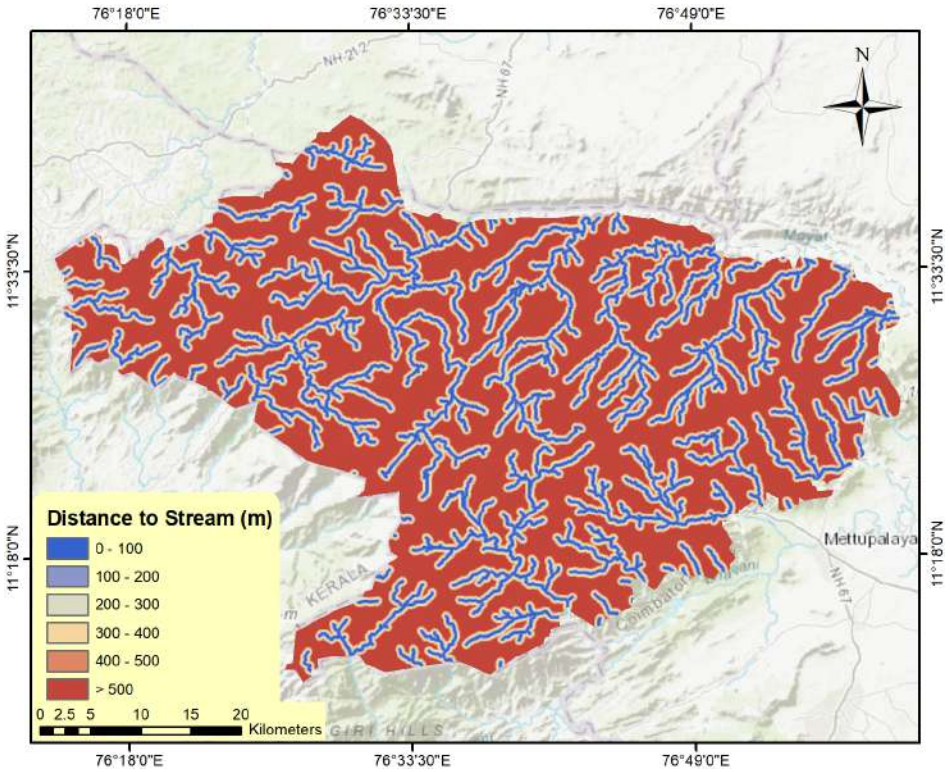
Distance to Lineament – Distance from geological fractures or faults that may control slope instability.



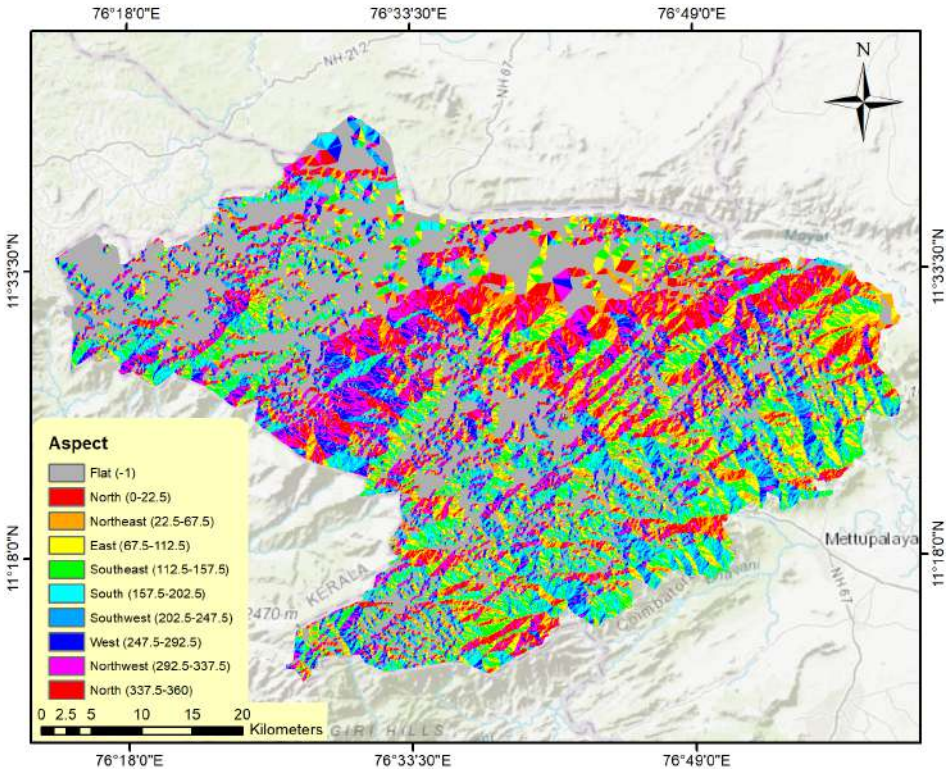
Distance to Road – Proximity to roads which can affect slope stability due to excavation or loading.



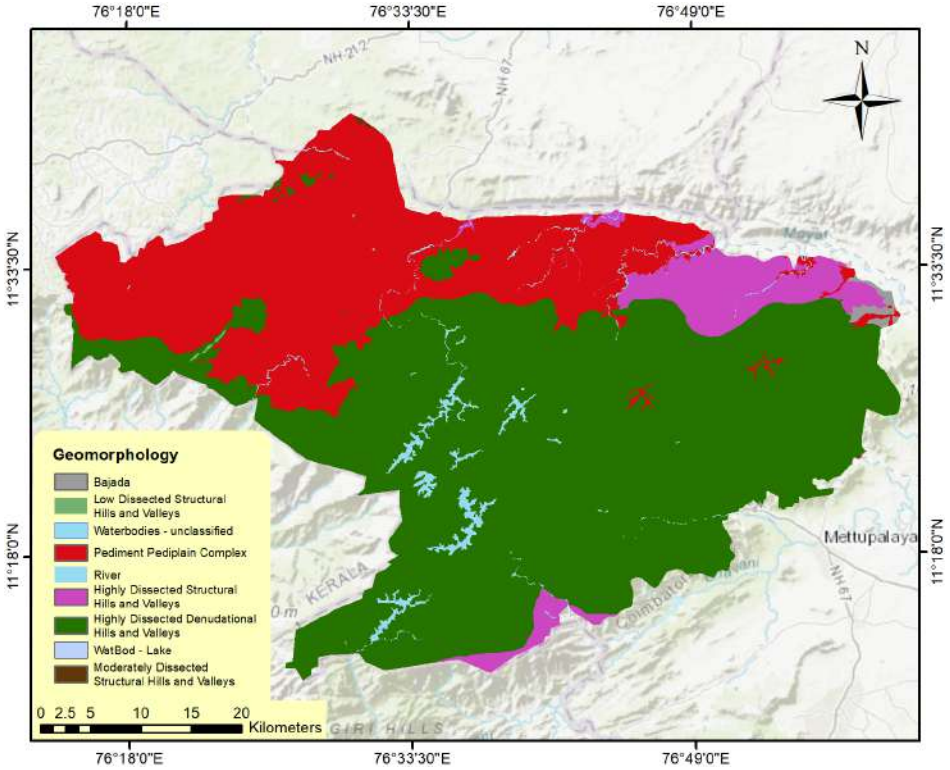
Distance to Stream – Distance from rivers or streams that can erode slopes and trigger landslides.



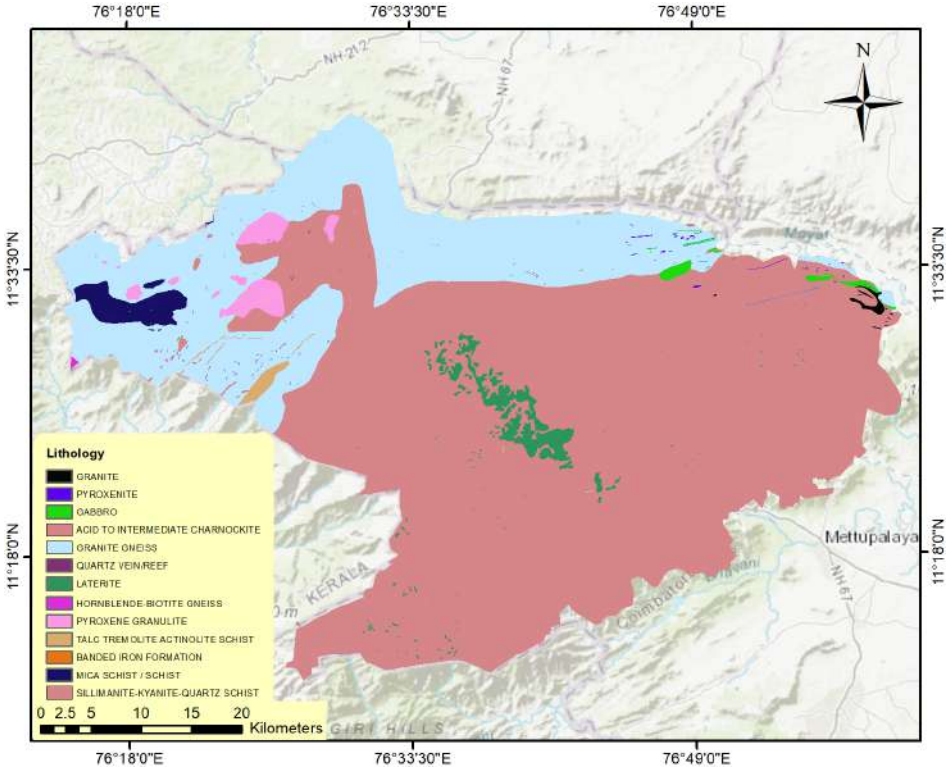
Aspect – Direction that each slope faces, influencing sunlight and water flow.



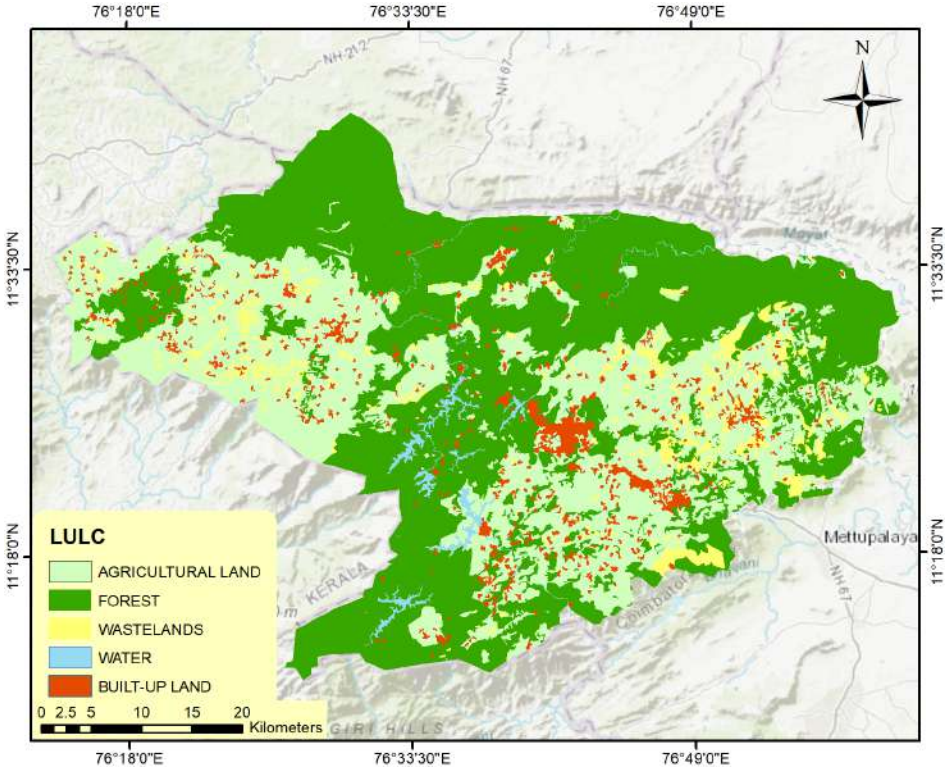
Geomorphology – Shape and form of the land surface that influences landslide susceptibility.



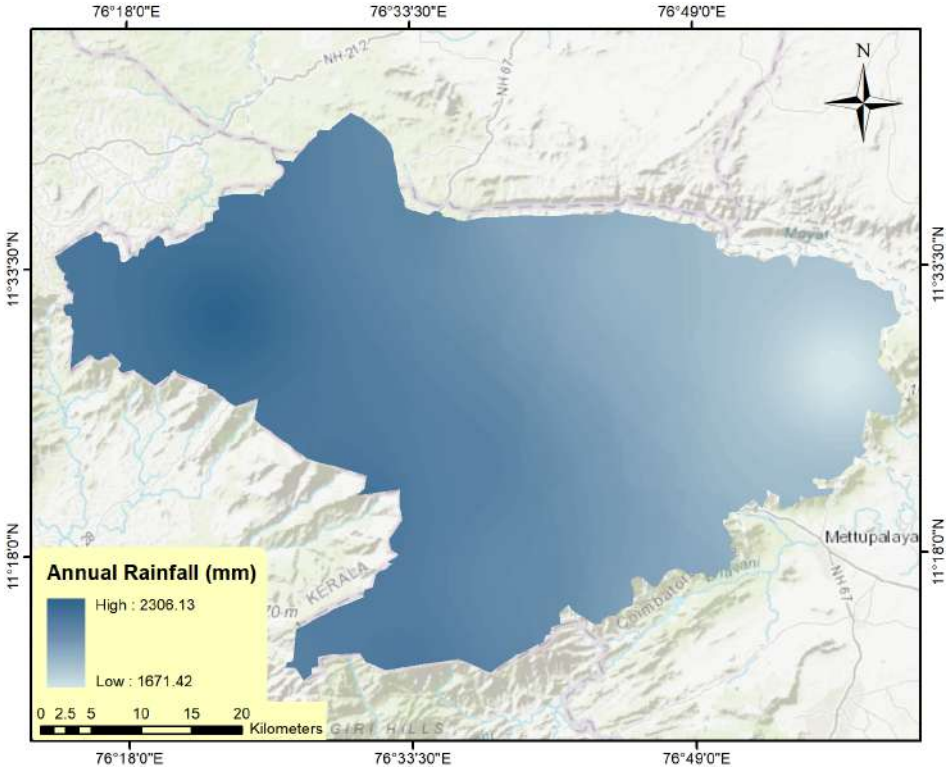
Lithology – Rock types and their properties affecting slope strength and stability.



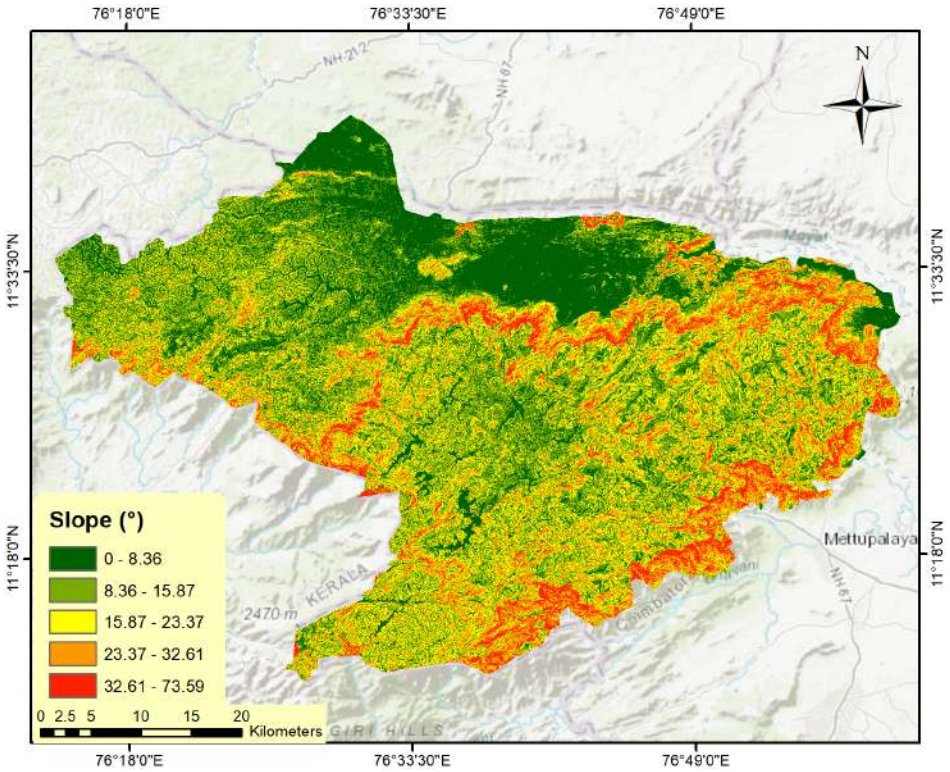
LULC (Land Use Land Cover) – Type of land cover such as forest, agriculture, or built-up areas.



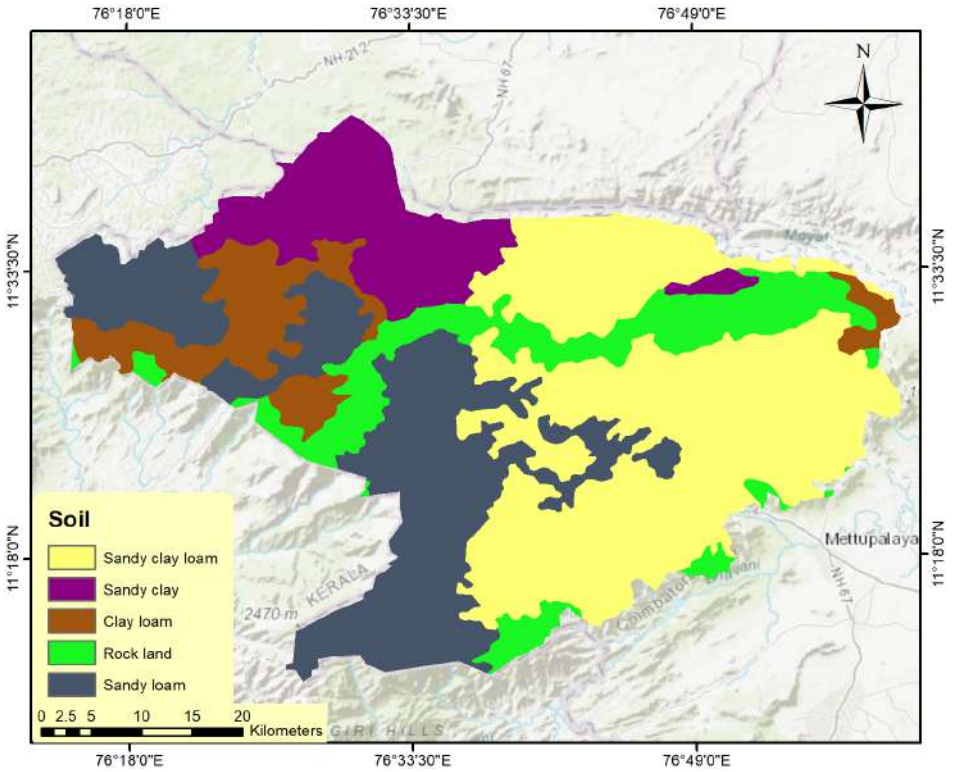
Rainfall – Annual precipitation, a major trigger for landslides.



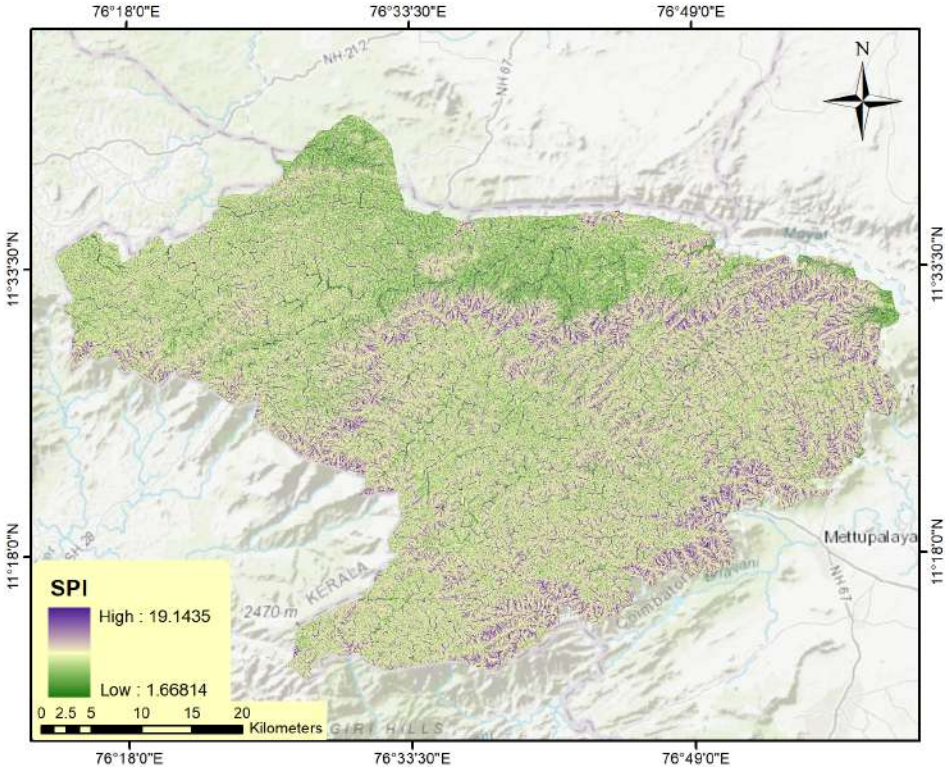
Slope – Steepness of the terrain, with steeper slopes being more prone to failure.



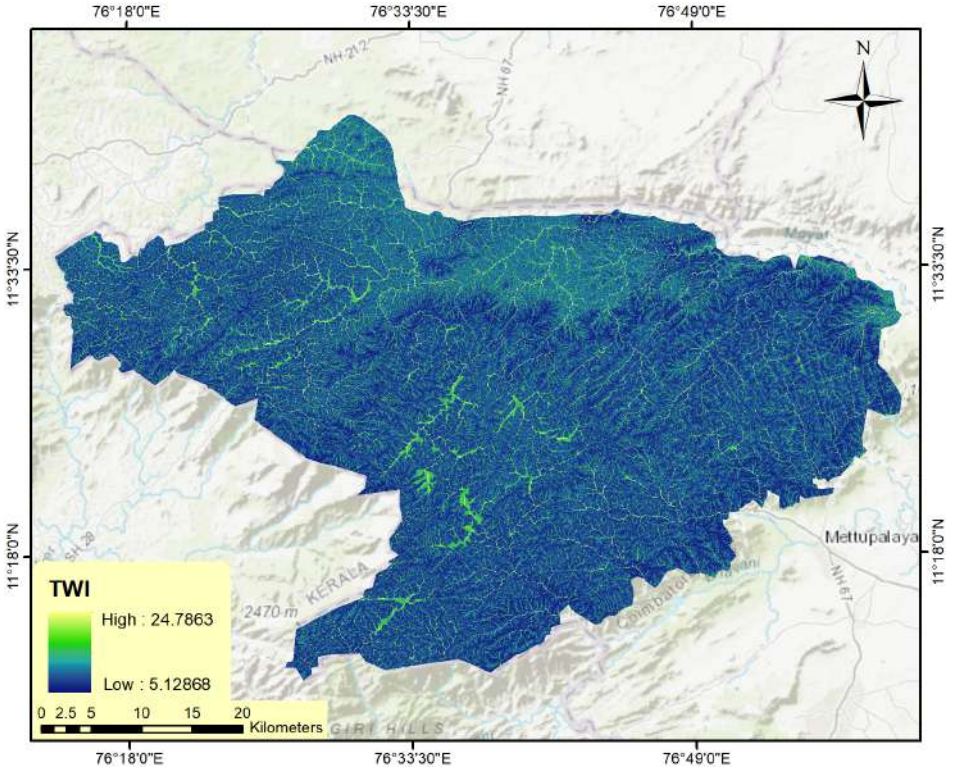
Soil – Soil type and texture, which control water infiltration and cohesion.



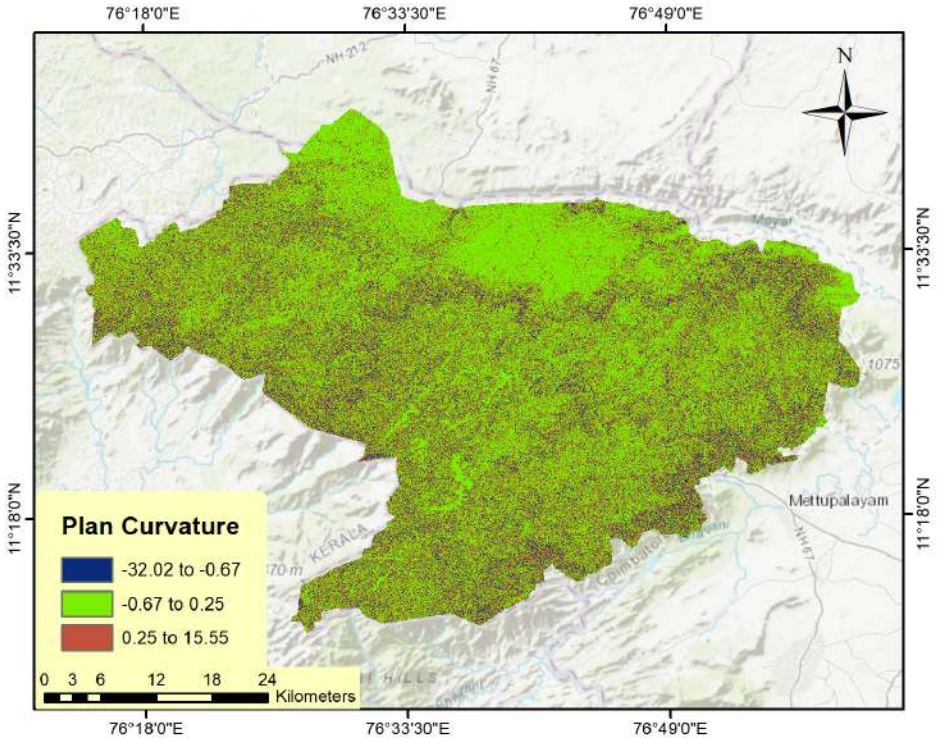
SPI (Stream Power Index) – Potential erosive power of surface water flow on slopes.



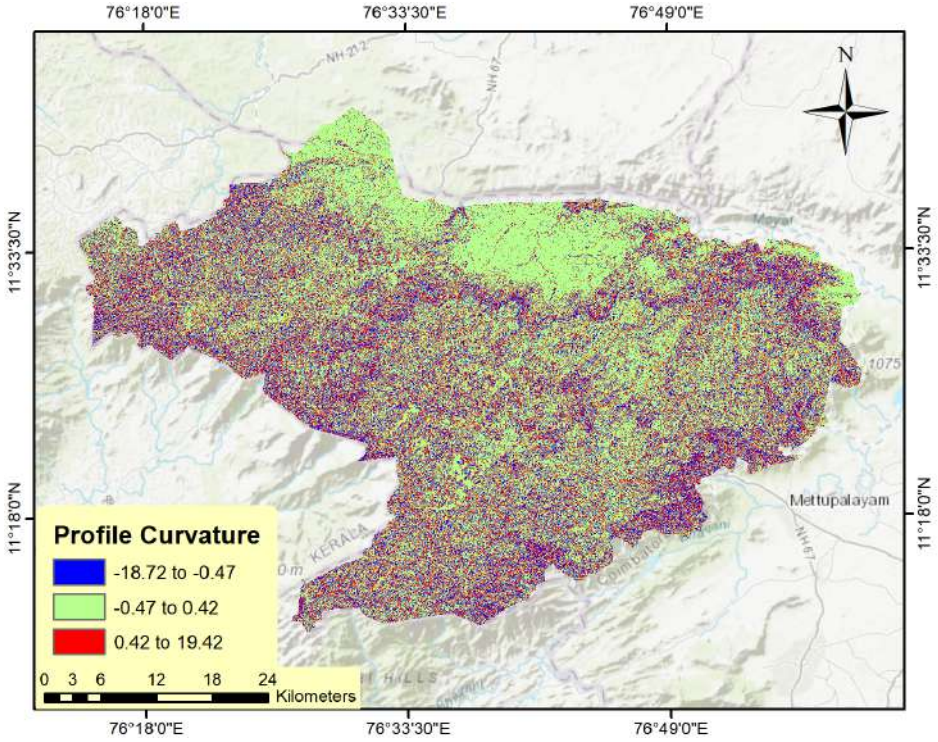
TWI (Topographic Wetness Index) – Likelihood of soil saturation based on terrain and drainage.



Plan Curvature – Curvature of slopes in horizontal plane, affecting water flow and erosion.



Profile Curvature – Curvature of slopes in vertical plane, influencing acceleration or deceleration of flow.



Data Preparation

- The data sourced from websites are truncated to a standard boundary of the Nilgiris District.
- DEM is used to create more features namely, elevation, slope, aspect, distance to stream, plan curvature, profile curvature, TWI, SPI
- A certain number of non-landslide points (value 0) are plotted as well and the data is merged with the landslide points (value 1).
- All data which has been in .tiff format for use in ArcGIS is converted to a singular .csv file for ease of handling using rasterio and pandas libraries in python.
- This .csv files is split into two – one with the model training data (with non-landslide and landslide points) and the other to make predictions on (without landslide point data).

Exploratory Data Analysis

This is a method to analyse the dataset to understand its main characteristics.

This involves summarizing data features, detecting patterns and uncovering relationships through visual and statistical techniques.

- Columns in Large Dataset:
  ['x', 'y', 'Aspect', 'Dist2Lineament', 'Dist2Road', 'Dist2Stream', 'Elevation', 'Geomorphology', 'Lithology', 'LULC', 'Plan_Cur', 'Pro_Cur', 'Roughness', 'Slope', 'Soil', 'SPI', 'TWI']
- Columns in Small Dataset:
  ['x', 'y', 'Aspect', 'Dist2Lineament', 'Dist2Road', 'Dist2Stream', 'Elevation', 'Geomorphology', 'Lithology', 'LULC', 'Plan_Cur', 'Pro_Cur', 'Roughness', 'Slope', 'Soil', 'SPI', 'TWI' , 'Final_Merged']
- Number of columns: - Large: 17 - Small: 18
  Number of rows: - Large: 2719466 - Small: 8118
- Total NaN cells in small: 0 - dtype: int64
  Total NaN cells in large: 0 - dtype: int64

EDA

Feature Correlation Matrix

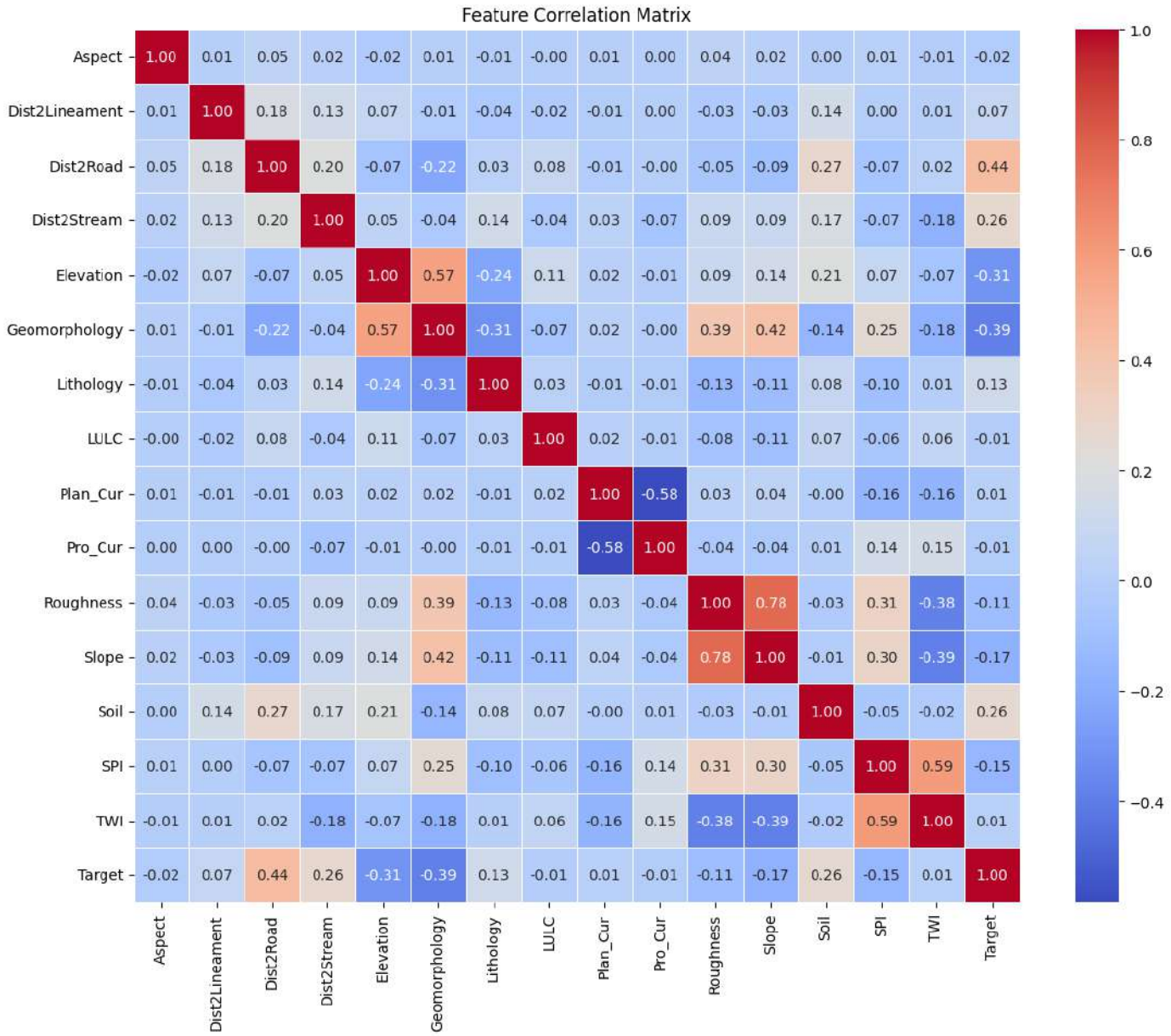
A feature correlation matrix assesses statistical relationships between variables.

It is used in feature selection to remove redundant (highly correlated) or irrelevant (poorly correlated with target) features.

Reducing correlation helps improve model performance by lowering noise, preventing overfitting, and decreasing training time.

Negative (inverse) correlation means one variable increases while the other decreases.

Key correlations observed: slope & roughness (+0.78), SPI & TWI (+0.59), elevation & geomorphology (+0.57), plan & profile curvature (-0.58); all below the 0.80 threshold, so acceptable.

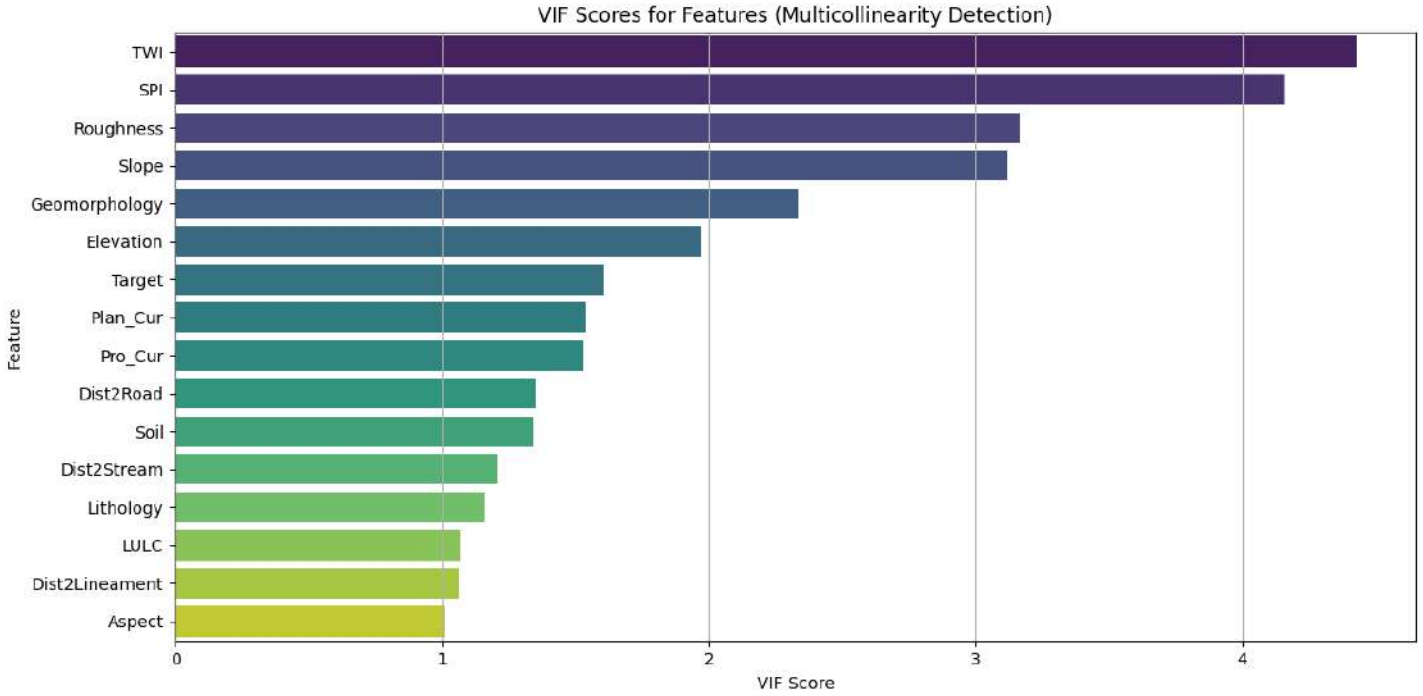


EDA

Value Inflation Factor

VIF (Variance Inflation Factor) measures how much a feature's variance is inflated due to multicollinearity with other features in a regression model.

The value inflation factor scores for the features are calculated as well to ensure that they do not exceed the VIF threshold of this study, which has been set as 5.

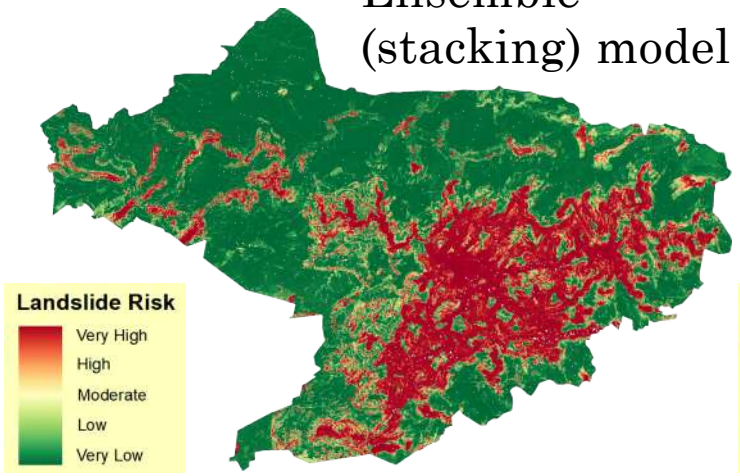


Initial Runs and Issues Faced

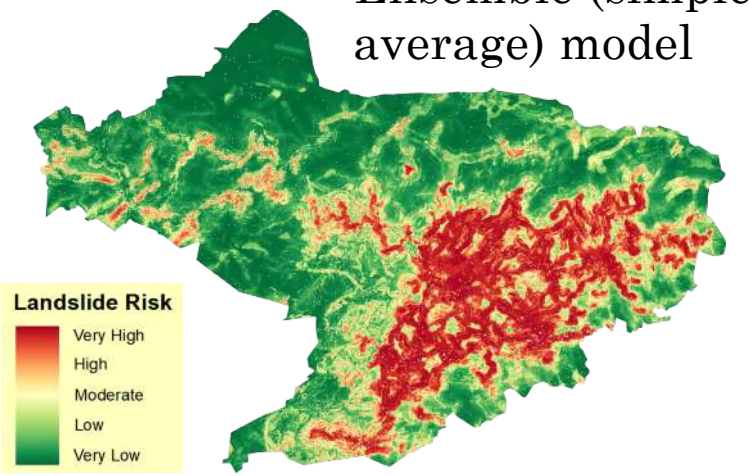
All models are giving very high feature importance to the feature distance to road.

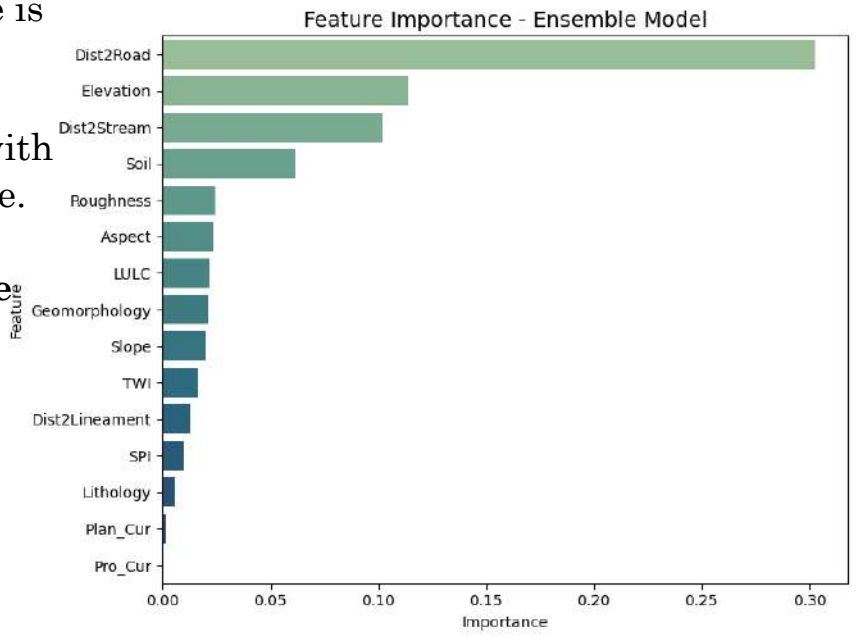
One potential cause is data leakage. The given feature can predict the target with very high confidence. **To generalize the results further we drop the feature.**

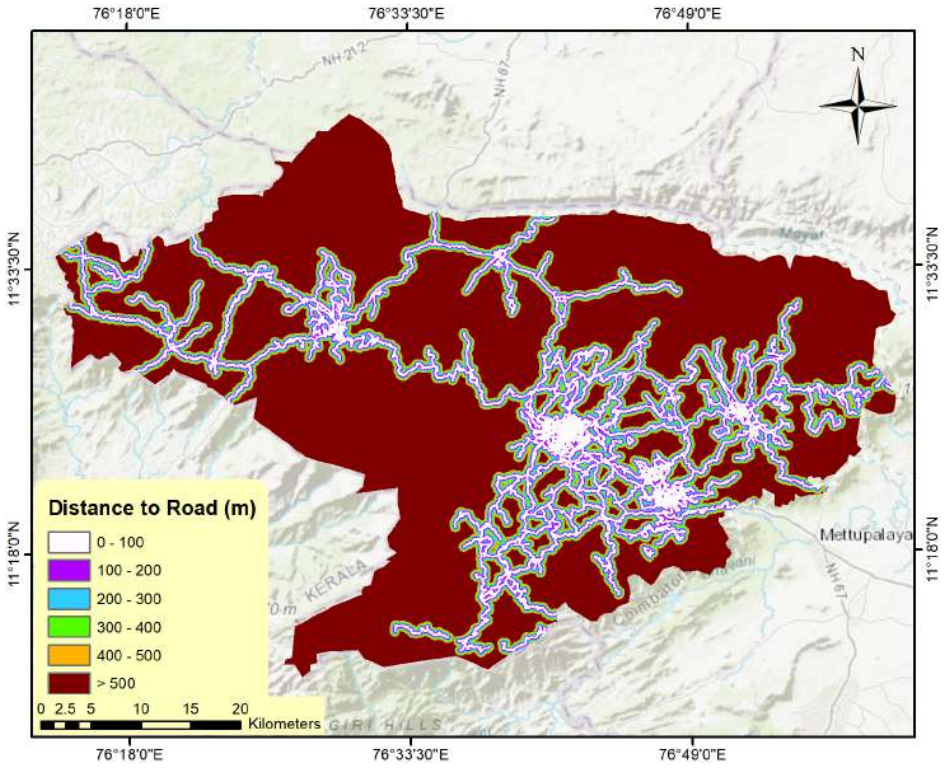
Ensemble (stacking) model



Ensemble (simple average) model





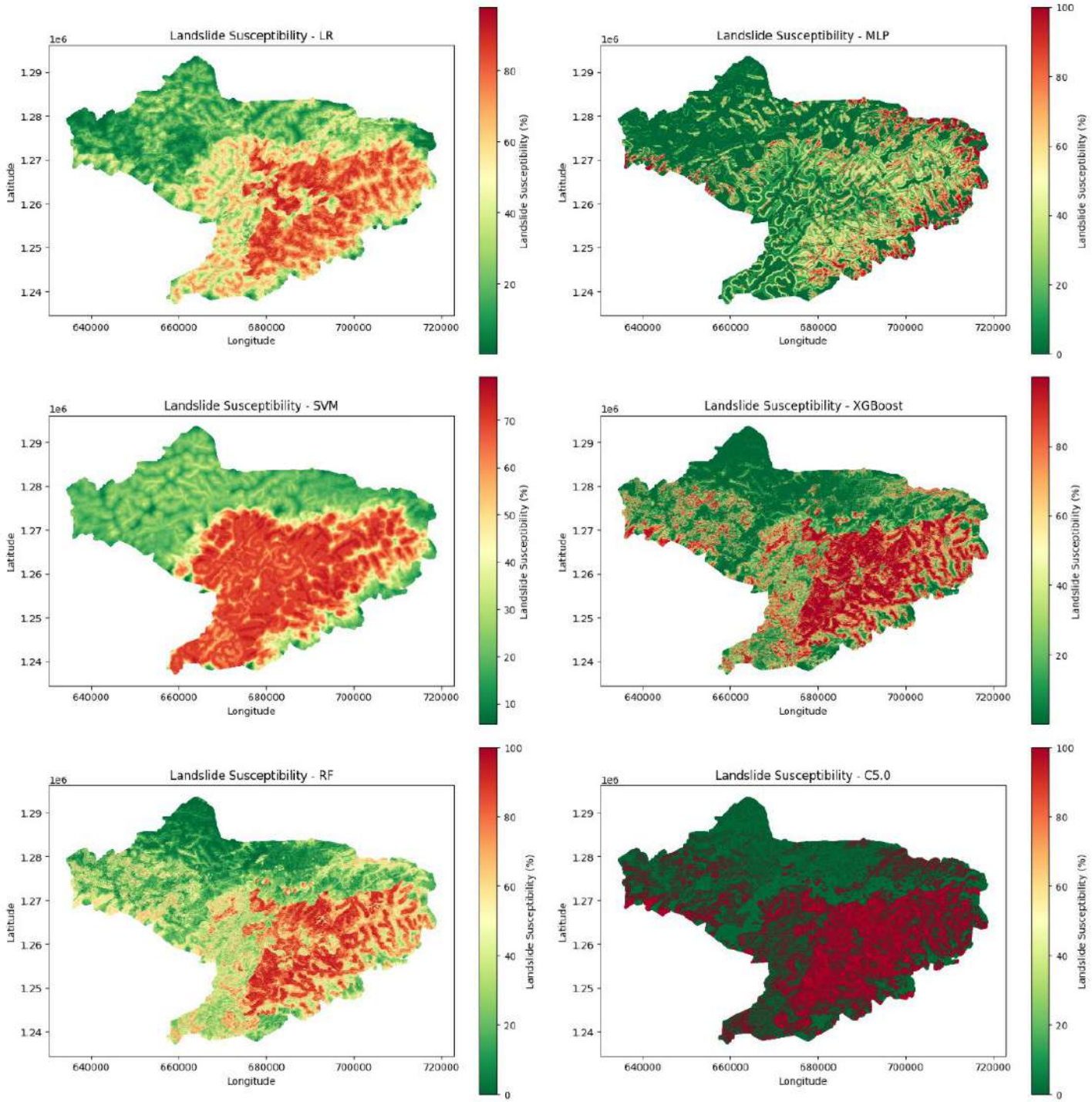


SVM model is overly reliant on the features elevation and distance to stream.

Cause found: Improper data preparation leading to negative infinity values not being dropped during trimming. This was corrected.

MLP model prediction patterns are very different in comparison to other models.

Cause not found hence CatBoost is used in place of MLP.



Models

LR (Logistic Regression) – Predicts susceptibility by estimating the probability of a landslide based on a weighted combination of input features.

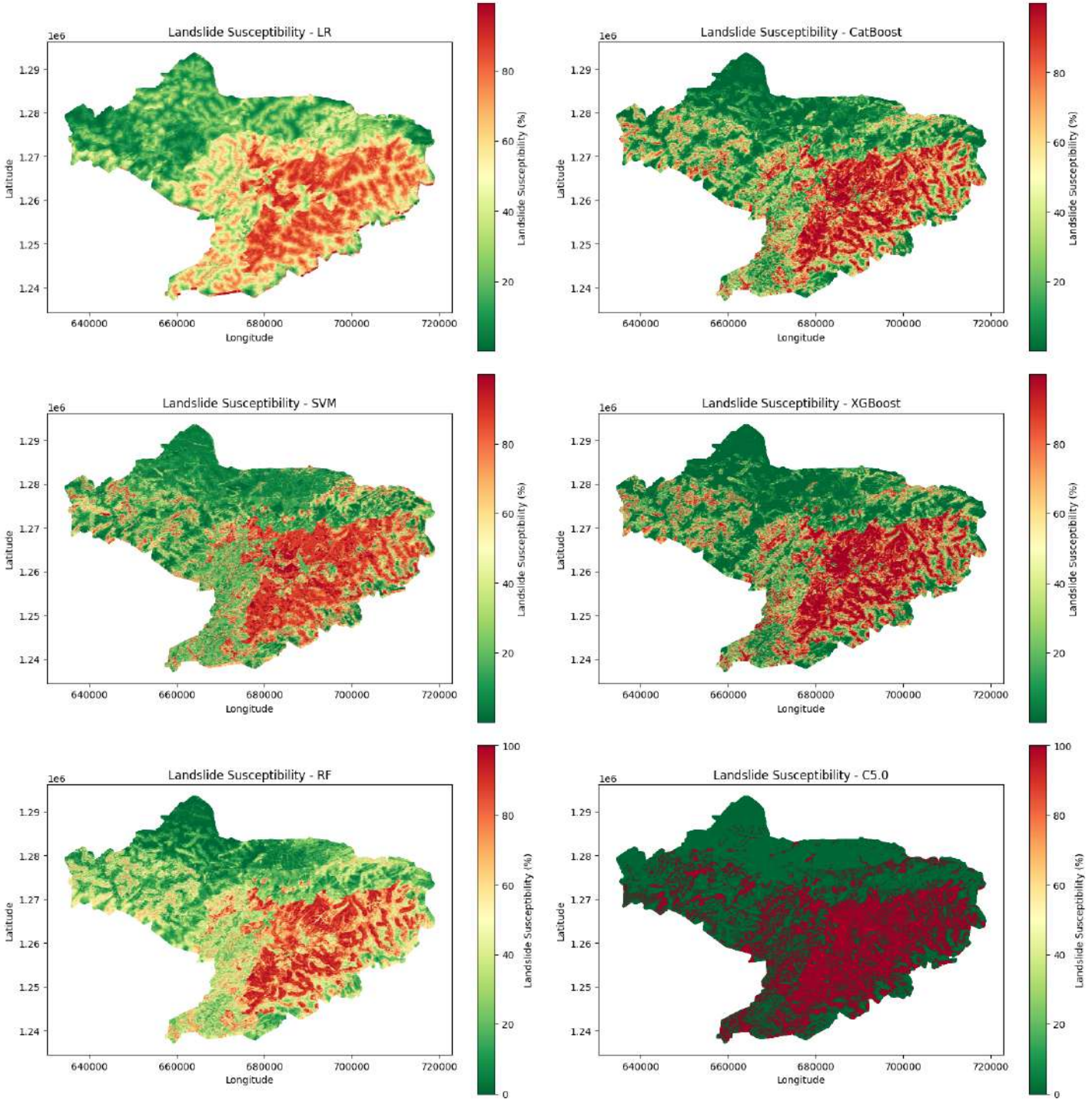
CatBoost – Uses gradient boosting on decision trees with categorical feature handling to model complex relationships.

SVM (Support Vector Machine) – Classifies areas by finding the optimal hyperplane that separates high- and low-susceptibility zones.

XGBoost – Applies gradient boosting with regularization to iteratively improve prediction accuracy on landslide-prone areas.

RF (Random Forest) – Uses an ensemble of decision trees to vote on susceptibility, reducing overfitting and improving stability.

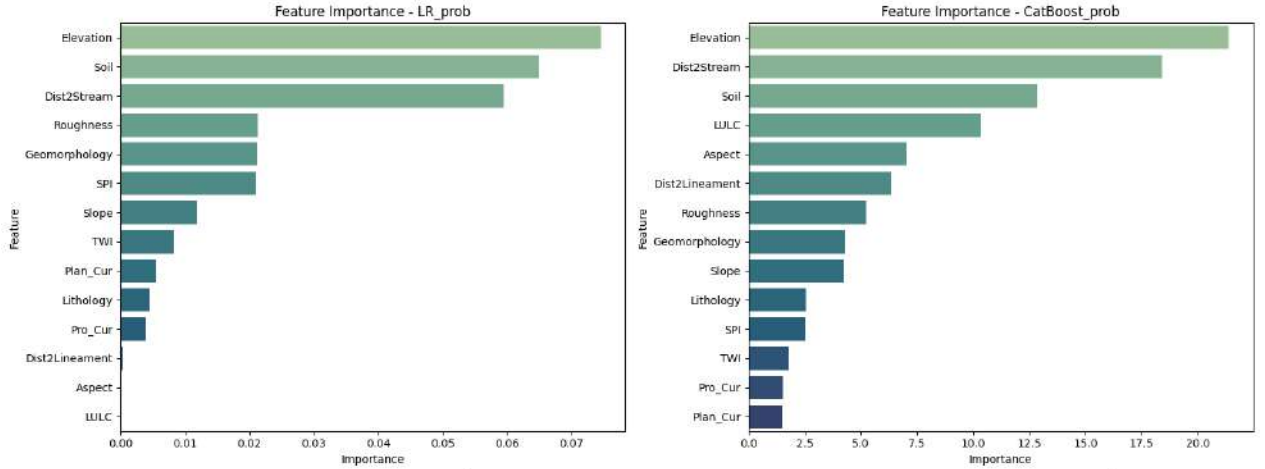
C5.0 – Builds a decision tree that splits features to classify areas into susceptibility categories based on information gain.

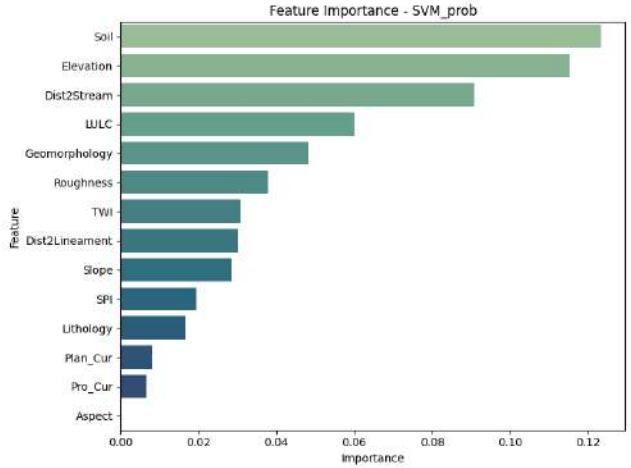


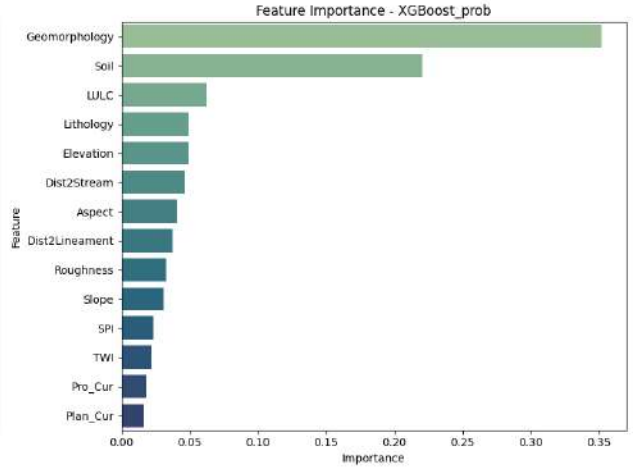
Logistic Regression Equation:

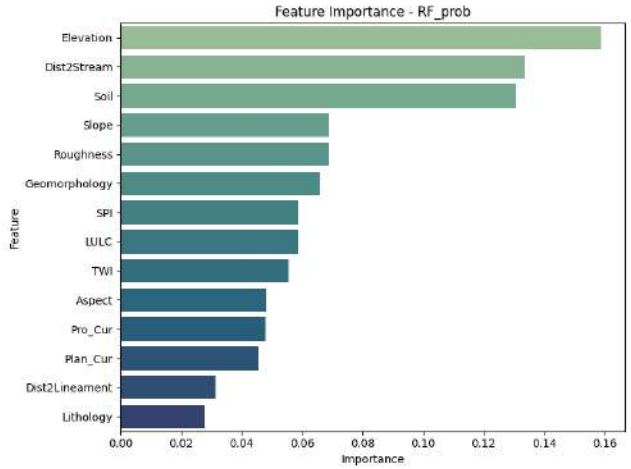
P(y=1) = 1 / (1 + exp(-(Aspect*(-0.0003) + Dist2Lineament*(-0.0002) + Dist2Road*(0.0008) + Dist2Stream*(0.0015) + Elevation*(-0.0014) + Geomorphology*(-0.0087) + Lithology*(0.0035) + LULC*(-0.0916) + Plan_Cur*(0.1483) + Pro_Cur*(0.1121) + Roughness*(0.0170) + Slope*(-0.0250) + Soil*(0.1689) + SPI*(-0.2830) + TWI*(0.2473) + 0.4900)))

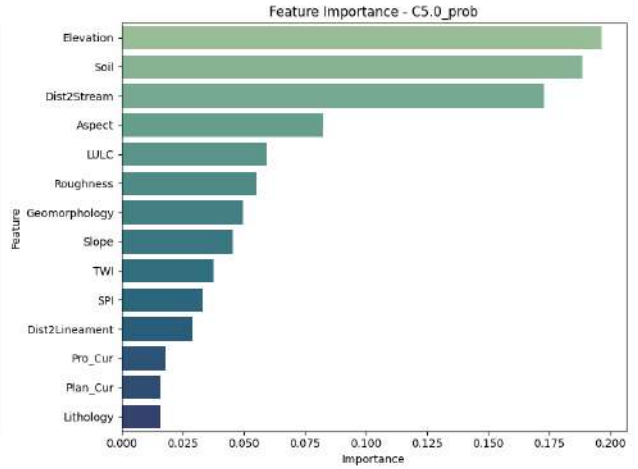
The rest of the models are non-linear hence cannot be expressed as an equation.





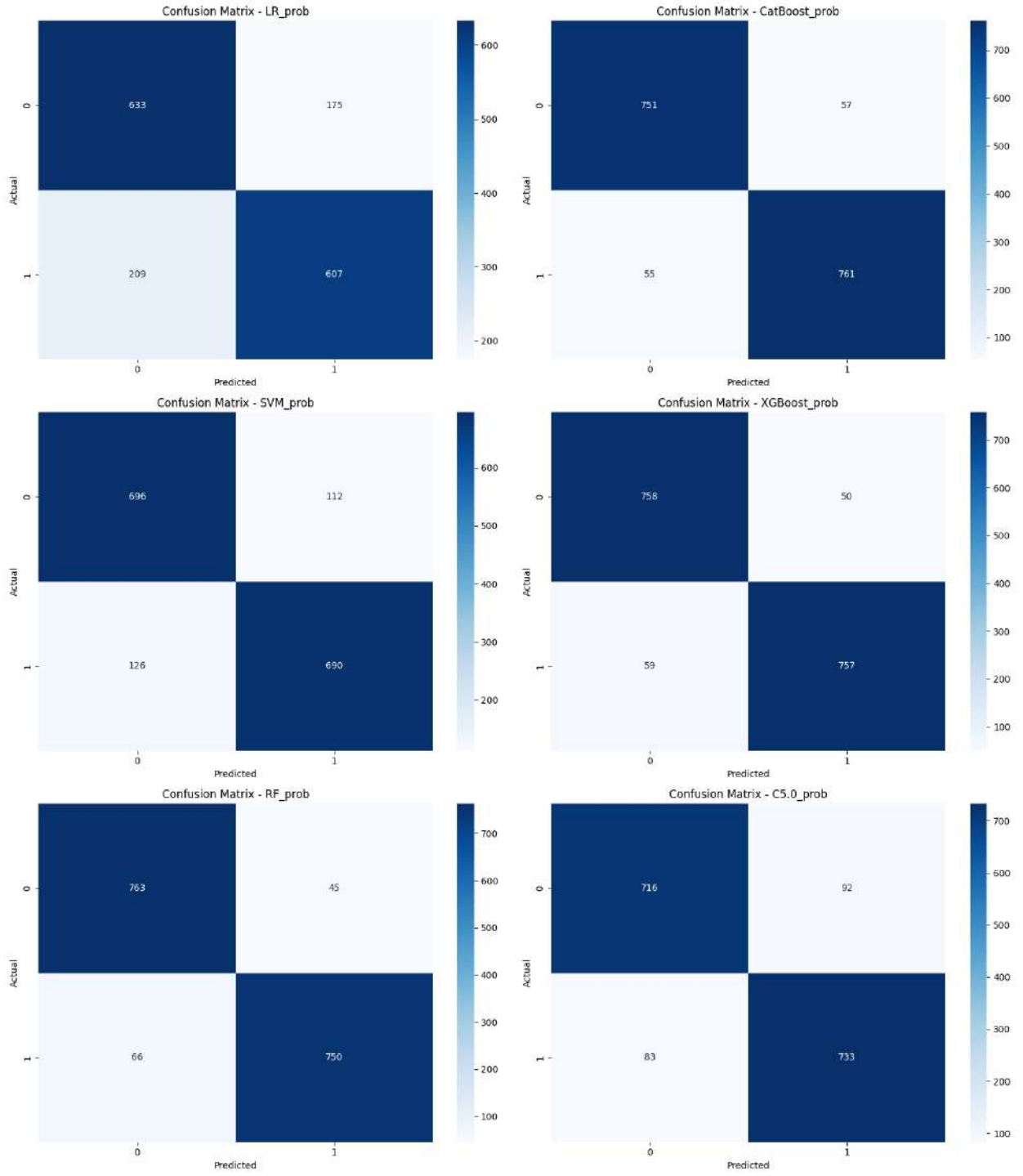






The confusion matrix visualizes the predictions based on true, false, positives and negatives. Higher numbers on the diagonal is preferable.

Logistic regression being the simplest model among all the models has relatively higher false positives and false negatives.



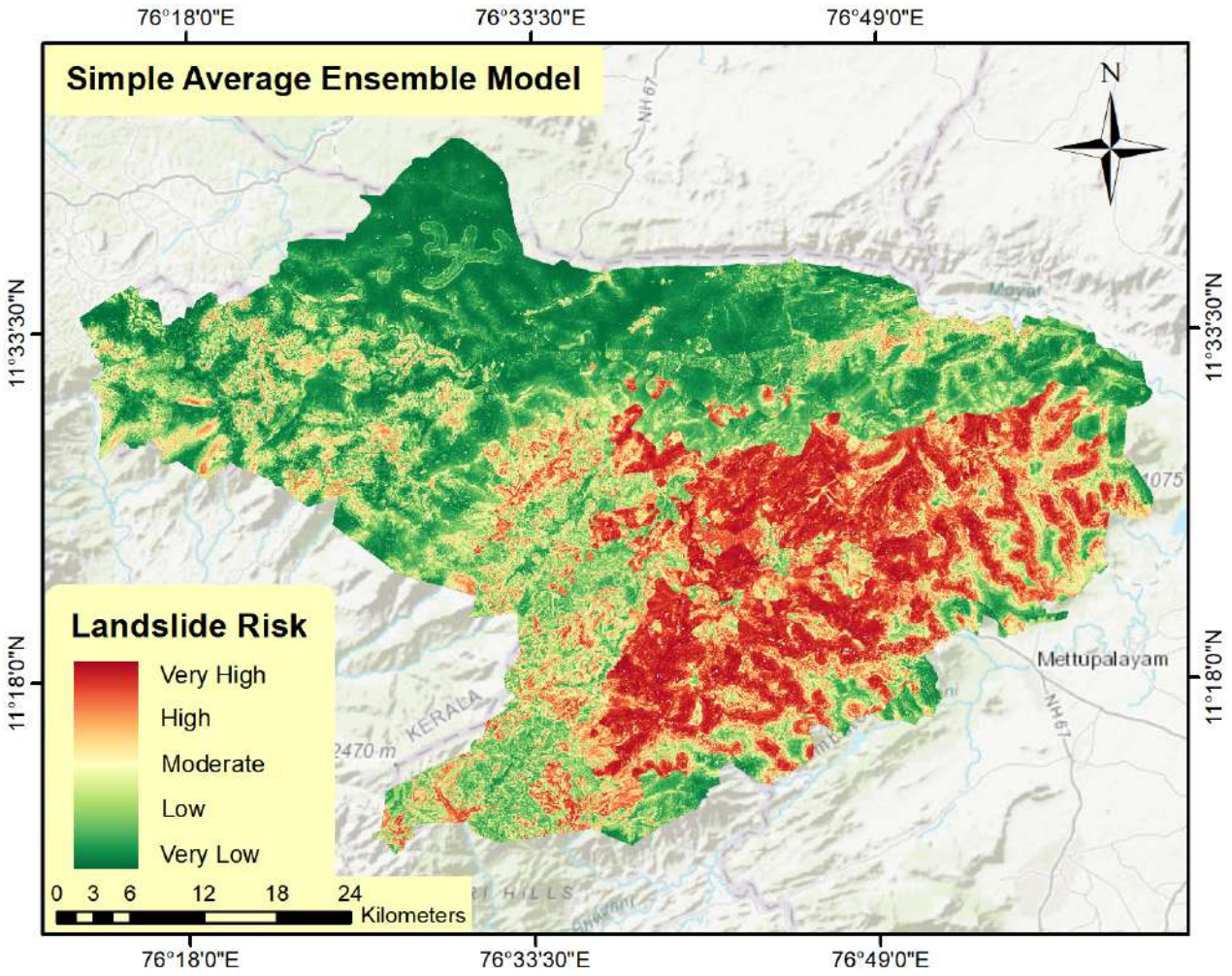
Results

Ensemble (Simple Average)

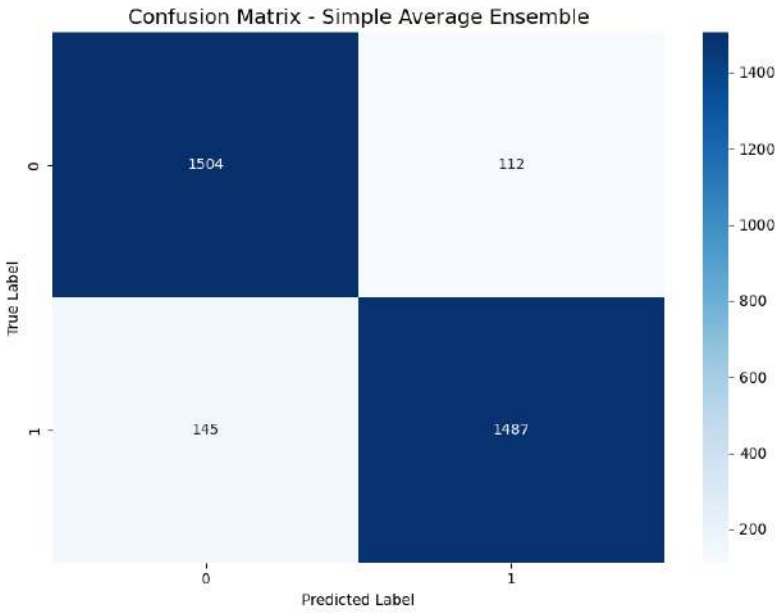
This method of ensembling averages the results obtained from all 6 models.

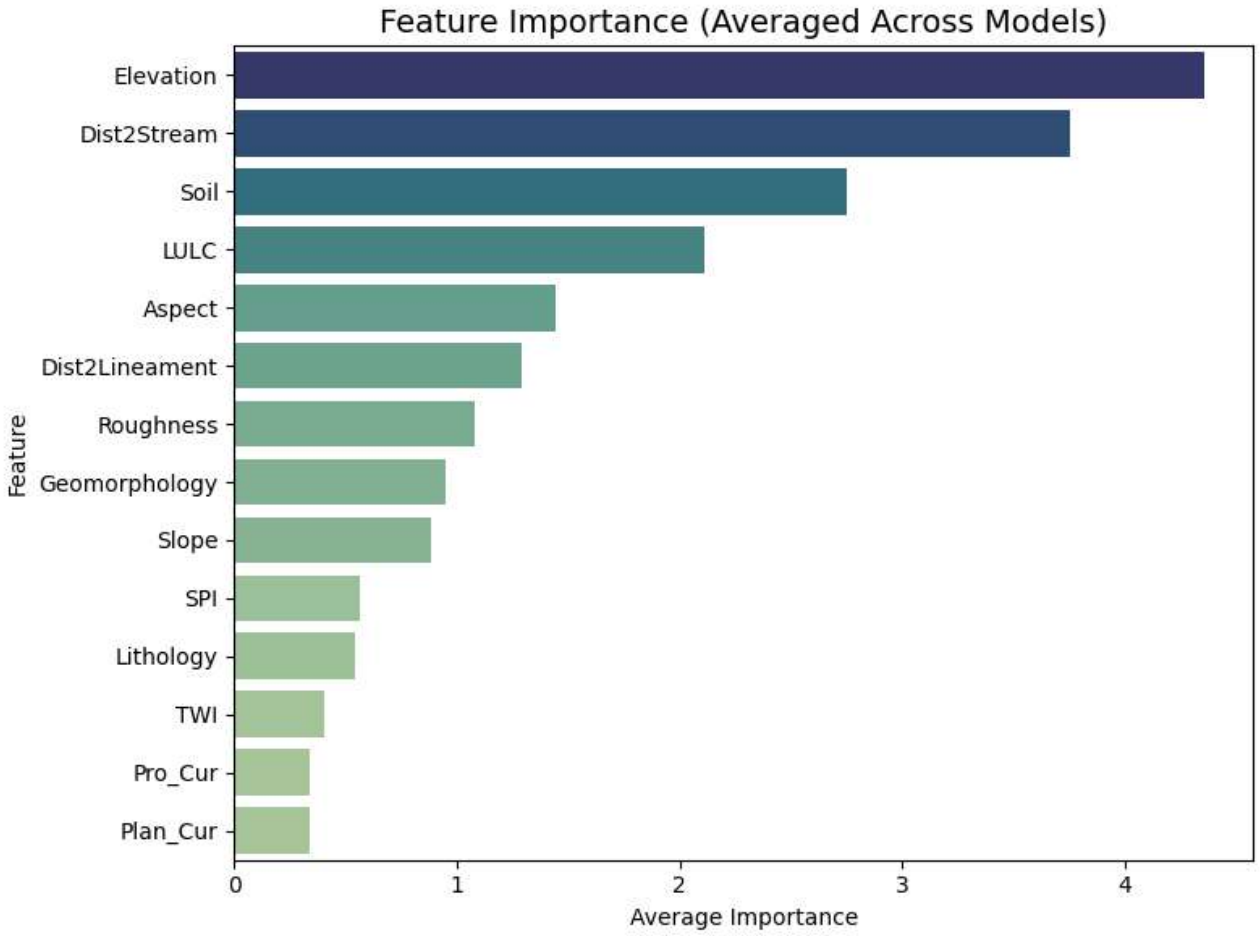
Equation for Ensemble (Simple Average):

Prob_avg(x) = (1/6) * [LR_prob(x) + CatBoost_prob(x) + SVM_prob(x) + XGBoost_prob(x) + RF_prob(x) + C5.0_prob(x)]



The top 3 features used commonly by all models to predict landslides in simple average ensembling method were elevation, distance to stream and soil.





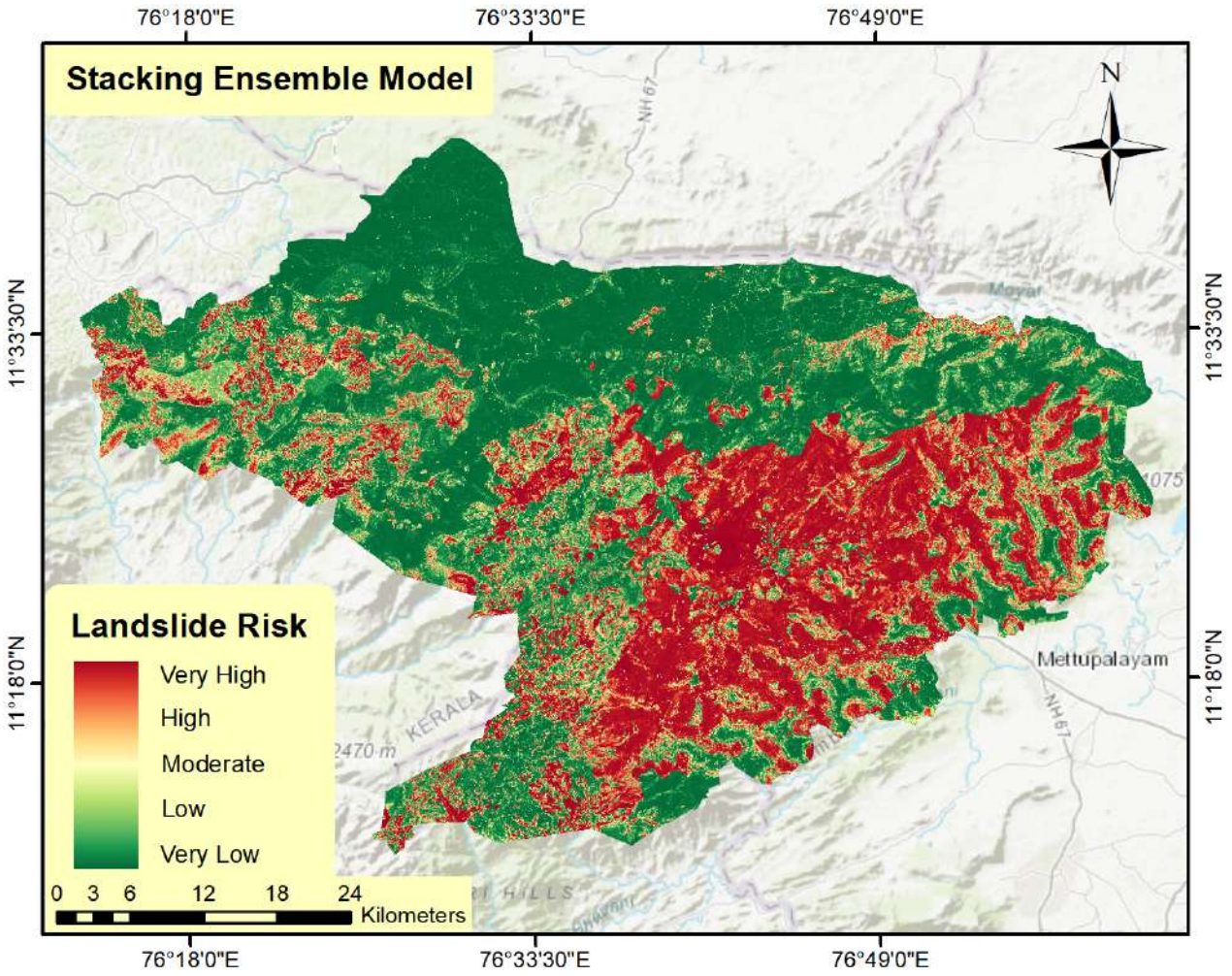
Results

Ensemble (Stacking)

Stacking is an ensemble learning technique where multiple base models are trained, and their predictions are used as inputs for a meta-model that makes the final prediction. It leverages the strengths of different models to improve overall accuracy.

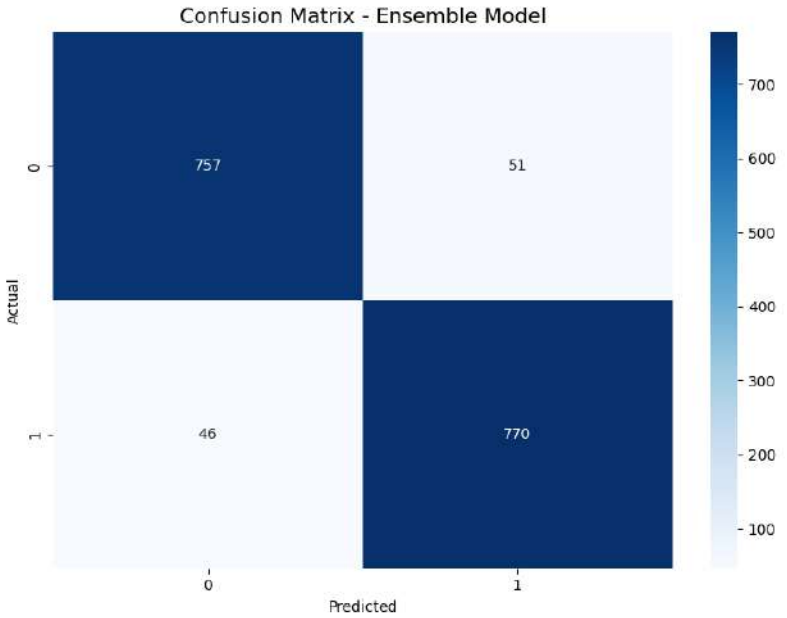
Equation for Ensemble (Stacking):

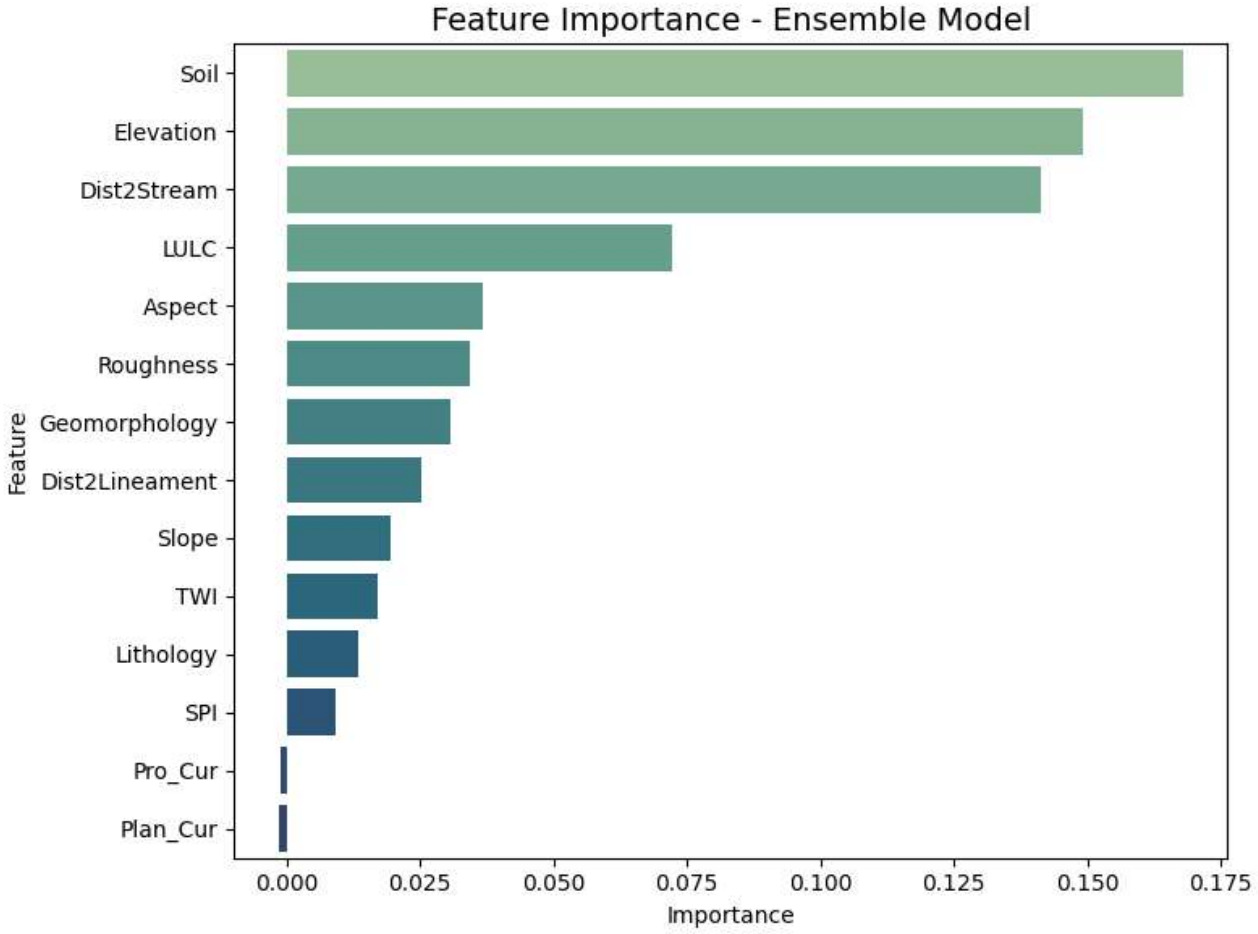
P(y=1) = 1 / (1 + exp(-(Aspect*(-0.1551) + Dist2Lineament*(2.5255) + Dist2Road*(0.9979) + Dist2Stream*(2.8599) + Elevation*(1.8058) + Geomorphology*(2.2595) + Lithology*(-0.0001) + LULC*(-0.0005) + Plan_Cur*(-0.0001) + Pro_Cur*(-0.0007) + Roughness*(0.0006) + Slope*(-0.0100) + Soil*(0.0107) + SPI*(-0.2832) + TWI*(-0.1188) + -0.1394)))



The top 3 features used commonly by all models to predict landslides in stacked ensembling method were soil, elevation and distance to stream again.

The order of top 3 features might have changed but the top 3 have repeated consistently in both cases.





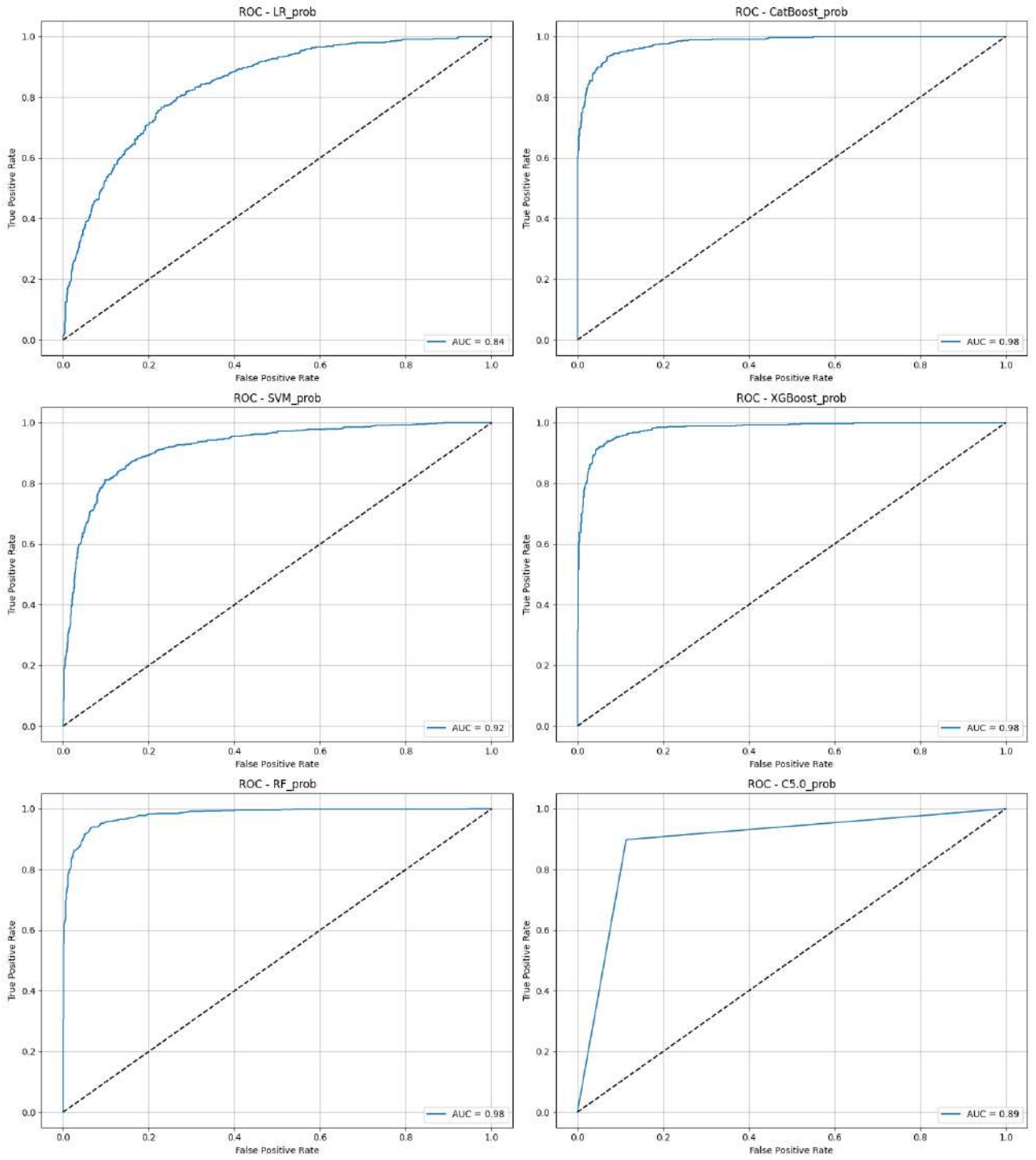
Validation of Results

The receiver operating characteristic (ROC) curve is used to evaluate the model's performance and overall accuracy.

The ROC plots the true positive rate (Y-axis) against the false positive rate (X-axis), with the ideal point at the top-left corner.

Model quality is indicated by the area under the curve (AUC), with values closer to 1 representing better performance.

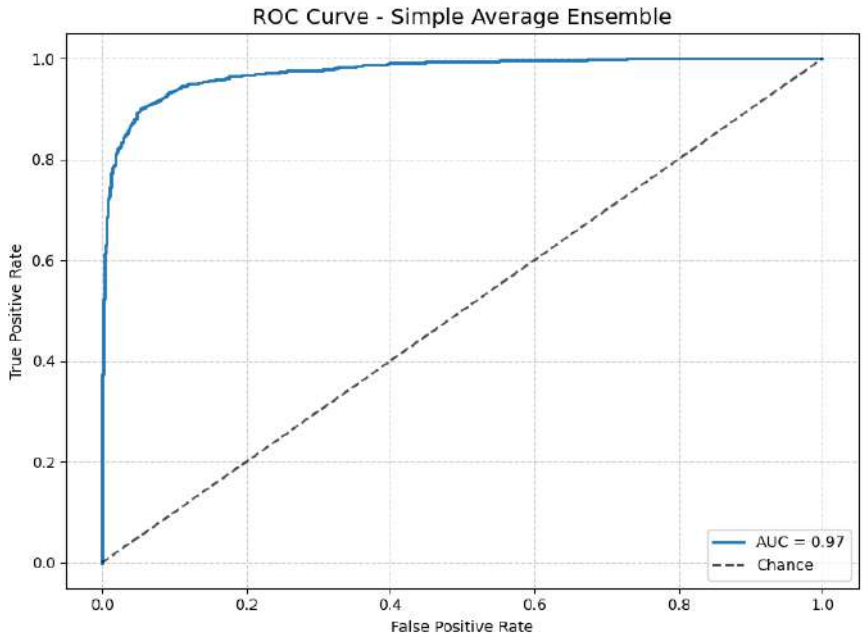
The ROC curve's AUC (area under the curve) value is divided into five categories as excellent (0.90–1.00), good (0.80–0.90), fair (0.70–0.80), poor (0.60–0.70), and fail (0.50–0.60) to ascertain the accuracy.

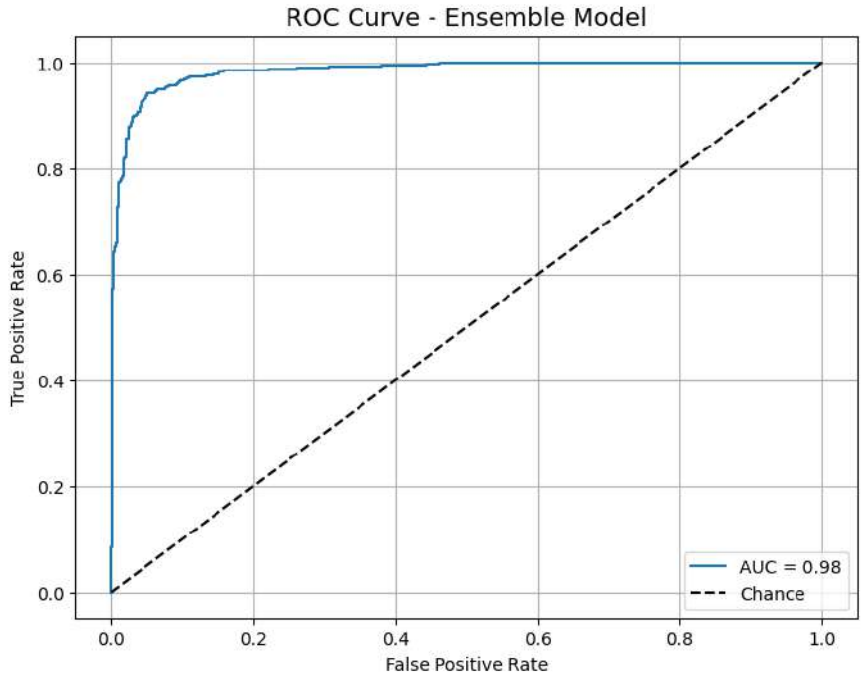


Validation of Results

The AUC values for the models used to predict landslide susceptibility of the Nilgiris district and their categories respectively are as follows:

Logistic Regression	0.825	Good
CatBoost	0.979	Excellent
Support Vector Machine	0.915	Excellent
XGBoost	0.979	Excellent
Random Forest	0.979	Excellent
C5.0	0.895	Excellent
Simple Average Ensemble Model	0.976	Excellent
Stacking Ensemble Model	0.985	Excellent





Conclusion

Landslide hazard mapping in the Nilgiris district is increasingly important due to rising climate change impacts.

The study uses the logistic regression, CatBoost, XGBoost, support vector machine, random forest, and C5.0. models, incorporating 15 causative parameters for landslide occurrence.

Ensembling of all 6 models is done using 2 methods, namely, stacking and simple averaging to produce the final results.

Both models provide clear, straightforward results, validated using the ROC curve.

The resulting zonation maps show spatial distribution of landslides but do not indicate timing, impact severity, or frequency.

These maps can guide government authorities in disaster prevention and urban planning by avoiding development in high-risk zones.